Untranscribed web audio for low resource speech recognition

Untranscribed web audio for low resource speech recognition

Andrea Carmantini, Peter Bell, Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

{a.carmantini, peter.bell, s.renals}@ed.ac.uk

Abstract

Speech recognition models are highly susceptible to mismatch in the acoustic and language domains between the training and the evaluation data. For low resource languages, it is difficult to obtain transcribed speech for target domains, while untranscribed data can be collected with minimal effort. Recently, a method applying lattice-free maximum mutual information (LF-MMI) to untranscribed data has been found to be effective for semi-supervised training. However, weaker initial models and domain mismatch can result in high deletion rates for the semi-supervised model. Therefore, we propose a method to force the base model to overgenerate possible transcriptions, relying on the ability of LF-MMI to deal with uncertainty. On data from the IARPA MATERIAL programme, our new semi-supervised method outperforms the standard semi-supervised method, yielding significant gains when adapting for mismatched bandwidth and domain.

Index Terms: speech recognition, semi-supervised training, domain adaptation, web data

1. Introduction

In recent years, interest in the practical use of automatic speech recognition has exploded thanks to voice user interfaces. Automatic speech recognition is now applied to disparate tasks, but poor generalization of models to unseen domains means that task-specific data is needed for optimal results. While obtaining speech data is often straightfoward, transcription is an expensive and time consuming task. Consequentially, for most languages it is impossible to find substantial amounts transcribed data for many (or all) domains.

The IARPA MATERIAL project aims to build models for the extraction of information from speech and text in low resource languages, combining speech recognition, machine translation, information retrieval and summarization to make multimedia sources accessible cross-lingually. For the speech recognition component, the most prominent challenge is to build a model for different domains without the use of any in-domain transcribed speech. Specifically, the training data provision consists of narrow-band conversational telephone speech while the test data includes wideband audio from news and entertainment broadcasts.

Domain adaptation is a well studied problem in automatic speech recognition. Multiple techniques have been developed for the adaptation of acoustic models, such as transfer learning [1, 2] and feature mapping [3, 4]. Semi-supervised and lightly supervised adaptation techniques use a base model trained on out-of-domain supervised data to generate targets on in-domain unsupervised data. The targets are then filtered or improved before being used in training as labels for the unsupervised data; this data is usually combined with the supervised data to train a new model [5].

In lightly supervised training, the unsupervised data used has non-verbatim, noisy transcriptions such as subtitles, and this additional information is used to refine the labels. Nguyen and Xiang [6] used subtitles to compute a biased language model used to decode the audio. Segments matching the subtitles were then aligned and selected by dynamic programming. Bell and Renals [7], constrained the decoding process using graphs generated from the relevant text.

Semi-supervised techniques have been developed to deal with the situation in which auxiliary text information is not available. Commonly, the best transcription hypothesis is used as a label sequence, and the unsupervised data is filtered at the frame, segment or utterance levels using model confidence metrics [5, 8]. Various techniques for confidence metric estimation have been developed, from the use of simple posterior probabilities to ensemble scoring or neural networks predictions based on engineered feature sets [9, 10]. Fraga-Silva et al [11], generated multiple hypotheses which were scaled by their posterior probability to produce soft targets. In conditions where data scarceness is a problem, the use of semi-supervised techniques has proven effective to adapt models to new domains.

Bandwidth adaptation seems to be a less researched problem. In mixed-bandwidth training [12, 13] the features from narrowband data are extended to the same dimensionality as the wideband data by adding zeroes, and models are trained on mixed data, learning information relevant to both bandwidths. Similarly in transform based methods different bandwidth data is mapped to the same feature space, relying on techniques such as constrained maximum likelihood linear regression [14]. Reconstruction methods, also called bandwidth extension, rely on statistical models to estimate the information from the missing frequencies [15, 16].

Recently, a new semi-supervised method based on the lattice-free maximum mutual information (LF-MMI) objective function was proposed [17]. In semi-supervised training based on LF-MMI, information is taken from a number of possible hypotheses and scaled by the probability of each hypothesis. The resulting soft targets incorporate information about model confidence and are especially helpful when using a weak decoding model where uncertainty is higher. This technique was applied to channel and bandwidth adaptation with good results [18].

In our work, we applied LF-MMI to simultaneous bandwidth and domain adaptation and extended the technique by forcing a more granular word lattice generation by promoting insertions. We present the LF-MMI objective, the semi-supervised technique and our extensions to this technique in section 2, the experimental setup in section 3, and experimental results in section 4.

2. Semi-supervised LF-MMI

Maximum mutual information estimation is a discriminative objective function that aims to minimize the conditional entropy of a class given an observation. When applied to speech recogni-

tion, the objective $\mathcal{F}_{MMI}$ is defined as

$$\mathcal{F}_{MMI} = \sum_{U} \log \frac{P(\mathbf{x}^u|\mathbb{M}^{num}_{\mathbf{w}^u})}{P(\mathbf{x}^u)} \quad (1)$$

$$P(\mathbf{x}^u) = \sum_{\mathbf{w}} P(\mathbf{x}^u|\mathbb{M}^{den}_{\mathbf{w}}), \quad (2)$$

where u is an utterance, $\mathbf{x}$ is a sequence of acoustic observations, $\mathbf{w}$ is a transcription, $\mathbb{M}^{num}$ is a graph containing the target transcriptions, and $\mathbb{M}^{den}_w$ is a graph containing all possible transcriptions.

The gradient is calculated as

$$\frac{\partial \mathcal{F}_{MMI}}{\partial \log P(x_t|j)} = \gamma^{num}_j(t) - \gamma^{den}_j(t) \quad (3)$$

Where $\gamma^{num}_j(t)$ and $\gamma^{den}_j(t)$ are the probability of being in state j at time t – the state occupancy probability, for the numerator and denominator respectively. Then the gradient with regard to the activations of a neural network used in a hybrid system is:

$$\frac{\partial \mathcal{F}_{MMI}}{\partial \log \mathrm{a}_t(s)} = (\gamma^{num}_j(t) - \gamma^{den}_j(t)) \frac{\partial \log P(x_t|j)}{\partial \mathrm{a}_t(s)} \quad (4)$$

where $\mathrm{a}_t(s)$ is activation of output s at time t.

A word level graph would make the calculation of the denominator computationally expensive, so in the LF-MMI implementation it is approximated by using a phone level graph.

For supervised training, $\mathbb{M}^{num}_{\mathbf{w}^u}$ is a graph containing alternative pronunciations of the correct transcription. For semi-supervised training, a lattice generated by the base model during decoding is used directly. This lattice is a weighted finite state transducer (WFST), a subset of the decode graph containing the states and transitions the beam search traversed; it can also be interpreted as a representation of the N-best transcriptions, determinized so that each word sequence appears only once [19].

The decode graph $HCLG$ is also a WFST:

$$HCLG = H \circ C \circ L \circ G\,, \quad (5)$$

where H is the structure of the HMMs representing phones, C is the context-dependency of the phones, L is the lexicon and G is the grammar or language model: the composed FST is a mapping from HMM states to words.

Sections of the generated decode lattice where the probability is distributed equally between a high number of paths, i.e. where the base model is uncertain, will have less effect on the semi-supervised model as they will result in low occupancy probabilities and smaller gradients. This means semi-supervised LF-MMI can deal directly with uncertainty.

2.1. Lattice generation biasing

In our work, we tested two methods to adjust the generation of lattices used for semi-supervised LF-MMI. We included a parameter ℓ, a bias to the transition weights $\mathbf{E}$ of the G WFST:

$$E^{bias}_{ij} = E_{ij} - \ell\,, \quad (6)$$

where in terms of an n-gram model of size N

$$E_{ij} = -\log P(w_t = j|w^{t-1}_{t-(N+1)} = i)\,. \quad (7)$$

We call ℓ the *insertion reward*, or negative insertion penalty; given two possible word sequences $\mathbf{w}_a$ and $\mathbf{w}_b$ of length n_a and n_b s.t. $P(\mathbf{w}_a|G) = P(\mathbf{w}_b|G)$ and $n_b > n_a$:

$$\log P(\mathbf{w}_a|G^{bias}) = \log P(\mathbf{w}_b|G^{bias}) - \ell(n_b - n_a)\,. \quad (8)$$

What this means in practice is that the model will prefer multiple short words to a long one, resulting in a proliferation of paths in the decoding lattice. While this could seem disadvantageous for training as the word lattice will have more uncertainty, when using LF-MMI the lattice is converted to phones and the probability of the states is marginalized over the paths; shorter words will result in higher granularity, thus representing better in which segments the model is most confident. Furthermore, insertions can be preferable in the context of this framework as a deletion would result in losing information regarding the underlying acoustics.

The second method involves scaling the likelihoods output by the acoustic model by an acoustic weight κ

$$P(\mathbf{w}|\mathbf{x}) = P(\mathbf{x}|\mathbf{w})^{\frac{1}{\kappa}} P(\mathbf{w}) \quad (9)$$

In practice, this is the same acoustic scaling factor used to correct for the assumption of independence between frames when using a non-discriminative objective [20]. Scaling the acoustic likelihoods when generating lattices means that the derived targets used for semi-supervised training will be more dependent on the acoustics than the language model, however a weak base acoustic model can result in a higher error rate.

3. Experiments

3.1. Experimental setup

We conducted experiments on the language packs distributed by IARPA for the MATERIAL programme. We use data from the Swahili, Tagalog and Somali releases. For each language the training set consists of 80 hours of conversational telephone speech (CTS) sampled at 8 KHz. The language packs also include two evaluation sets, one for narrowband development and one for the domains targeted by the programme. The first one contains 20 hours of speech in the same domain and bandwidth as the training set, i.e narrowband CTS. The second one consists of 20 hours of wideband recordings taken from broadcasts and web media, mostly news and entertainment, and was used as our adaptation target.

Furthermore, for our unsupervised experiments on Somali we collected 600 hours of additional unlabeled audio data from the web. 300 hours of recordings were taken from Voice of America Somali, a website that produces radio programs and videos on world news and another 300 hours were obtained from YouTube videos in Somali. This unlabeled dataset is wideband and covers similar domains to those present in the adaptation evaluation set.

The acoustic models in our experiments all share the same architecture, consisting of 12 factored time-delay neural network (TDNN-F) [21] layers with 1024 units and a 128 dimensional linear bottleneck, preceded and followed by one fully connected layer with 1024 units. Factored TDNN layers are 1 dimensional convolutions on the time axis where the parameters matrix is factorized through singular value decomposition into two semi-orthogonal matrices; the linear bottleneck is the dimension shared by the two factor matrices.

All layers use batch normalization and the ReLU activation function. The models are trained using natural gradient stochastic gradient descent with a LF-MMI objective, using cross entropy as a secondary objective for regularization [22]. We use a

Table 1: *Baseline results (WER %) on the CTS and wideband (WB) evaluation sets for models trained on 80h of CTS data for the three MATERIAL languages. In these experiments the WB dataset was downsampled.*

Model	CTS	WB
Swahili	39.4	44.6
Tagalog	42.6	45.8
Somali	55.5	61.6

dropout schedule where the dropout probability starts growing linearly from 0 to 0.5 after seeing 20% of the data, then decreases linearly after seeing 50% of the data. Each model was trained for 6 epochs.

For each of the three languages we trained a trigram language model on about 300 million words scraped from the web. The models employ KneserNey smoothing and backoff, and use a limited vocabulary made up of the 150 thousand most common words in the scraped text data. Pronunciations for these words were generated using a grapheme to phoneme model trained on manually compiled lexicons released by IARPA with the language pack. The text used to optimize interpolation hyperparameters for these language models was taken from a domain similar to our wideband dataset. Our language models are thus adapted to our target domain.

We used two separate sets of features for narrowband and wideband data. For narrowband data, we use 24 mel bins between 125 and 3800 Hz to compute filterbank features, then add pitch frequency and probability of voicing. For wideband data, we use 40 mel bins between 0 and 8000 Hz to compute filterbank features.

Table 1 reports baseline results for our model on the three MATERIAL languages. For these results, no wideband adaptation was used, so the wideband data was downsampled to use the same features as the CTS data. Our results compare favourably to previously published results on the same dataset [21, 23].

3.2. Feature-based adaptation

Semi-supervised techniques are largely influenced by the base model used for decoding the unsupervised data. To obtain a better base model, we experimented with 4 feature-based adaptation methods: mean and variance normalization, iVectors [24, 25], xVectors [26], and multilingual bottleneck features [27].

We trained an universal background model on the 80 hours of CTS data and we used it to extract 100-dimension iVectors. Both in training and evaluation, the iVectors were computed online, with a new one computed for every 10 frames of speech. As such, the iVectors capture more local variability and are less focused on speaker information.

We compared the iVectors with xVectors extracted using a network trained on data from Switchboard, NIST SRE and VoxCeleb data. We reduced the dimensionality of our xVectors to 100 through principal component analysis. The network was trained mostly on narrowband CTS, but the presence of the data taken from VoxCeleb, closer to our target domain, could bring useful information for the adapted model. Full details about the xVector network are in [26].

Similarly to xVectors, our aim in extracting multilingual bottleneck features is to aid the robustness of the model through the transfer of knowledge from other languages. The bottleneck network was trained using 23 language packs from the IARPA BABEL challenge, using 80 hours of CTS data for each language in a multilingual setup. The network architecture comprised 7 TDNN layers with 4096 units, a fully connected bottleck layer with 78 units and a final fully connected layer with 4096 units. Bottleneck features were extracted for the Somali data and used as auxiliary features to the filterbank inputs.

Table 2: *Somali results (WER %) for feature-based adaptation methods. All models are trained on 80h of CTS data only. In these experiments the WB dataset was downsampled.*

Model	CTS	WB
Baseline	55.5	61.6
• MVN	55.9	59.1
• iVectors	54.4	63.1
• xVectors	54.3	61.4
• mBNF + CMVN	**53.7**	**57.7**

Table 2 shows the effect of mean and variance normalization and auxiliary features for the two Somali evaluation sets. Mean and variance normalization gives us better results on the target domain for a small performance hit on the CTS set. In contrast, iVectors improve results on the matched data while being deleterious for the out-of-domain data. We believe this is due to MVN being data agnostic, while the iVectors are dependent on the data used in training the UBM. Using CTS data only, our iVectors do not seem able to capture well the variability present in the target dataset.

Both sets of features we trained on external data for knowledge transfer, xVectors and multilingual bottlenecks, gave us gains on both in- and out-of-domain data over the baseline, with xVectors being poorer than MVN for adapting to the target data. Using multilingual bottleneck features gave us the best results overall for both in- and out-of-domain data. As the bottleneck network was trained only on CTS data from the 23 languages used, the improvement on the out-of-domain dataset must be due to better linguistic information extraction and not to domain adaptation per se.

3.3. Semi-supervised LF-MMI

We chose the model using bottleneck features as the base to generate the targets for semi-supervised training. To compute the features needed as input to this model, we downsampled the unsupervised wideband data to 8 KHz. A new model was then trained interleaving the supervised and unsupervised data, halving the learning rate when training on the unsupervised data.

During lattice generation, we tested different insertion rewards and acoustic weights. We generated lattices from graphs biased with different insertion rewards, then used the targets computed from the lattices to train separate semi-supervised models. Figure 1 shows the relative change in word error rate when the targets used to train the semi-supervised model were extracted from these lattices. Result are shown for the 300 hours of unsupervised data taken from Voice of America and for the 300 hours taken from YouTube. For both subset of the data, we found that a reward of 1 gave us the best results.

Figure 2 shows the relative change when the lattices were biased by different acoustic weights. In this case, while a weight of 1.3 helped for the Voice of America data, the models trained on YouTube data degraded when assigning higher weights to

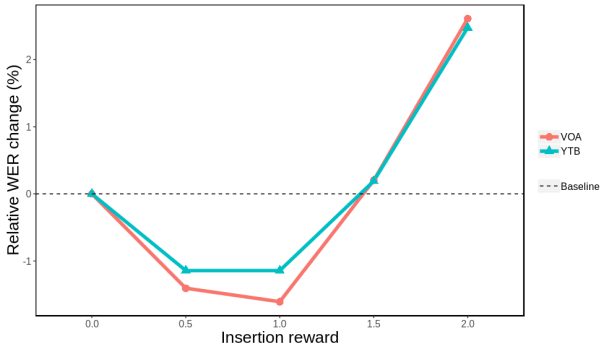

Figure 1: *Relative change in WER for semi-supervised models when using different insertion rewards in generating lattices. VOA and YTB refer to the Voice of America and YouTube subsets of unsupervised data.*

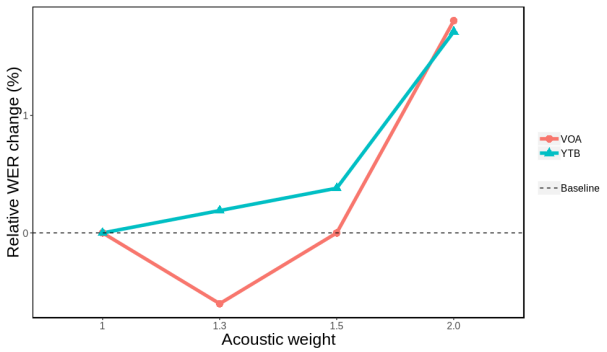

Figure 2: *Relative change in WER for semi-supervised models when using different acoustic weights in generating lattices. VOA and YTB refer to the Voice of America and YouTube subsets of unsupervised data.*

the acoustic model. We believe this difference is due to the noisiness of the YouTube data collected resulting in less reliable acoustic probabilities.

Our semi-supervised models are adapted to the new domain but still based on narrowband features. To adapt our model to the higher bandwidth, we made use of the targets generated for the unsupervised data and paired them to the corresponding wideband features. This let us train a new model on the unsupervised data only, without needing the narrowband CTS data.

We trained semi-supervised and unsupervised models on targets from biased and unbiased lattices. The biased lattices are generated using an insertion penalty of 1, as this is the condition that gave us stable gains across the two data subsets.

Our results are in table 3. The semi-supervised training on downsampled unsupervised data gave us great improvements on the out-of-domain evaluation set, with a 10.7% relative WER reduction over the baseline and an additional improvement of 0.5% when using biased lattices.

Surprisingly, our results got better when using only the unsupervised data paired with the relative wideband features, with an improvement of 14% relative WER, 3.7% more than the semi-supervised model. This shows that in this specific case the wideband features contain more relevant information than the combination of narrowband and bottleneck features. Also, it

Table 3: *Semi-supervised results (WER %) on Somali. The unsupervised models were trained on 600 hours of unlabeled data. The semi-supervised models were additionally trained using 80 hours of labeled CTS data, downsampling the wideband data.*

Model	**CTS**	**WB**
mBNF + CMVN	53.7	57.7
mBNF semi-supervised	56.6	51.5
mBNF semi-supervised + Ins. Rew. 1	54.2	51.2
Unsup. (16k)	-	49.6
Unsup. (16k) + Ins. Rew. 1	-	**48.9**

demonstrates that this model can learn well enough from noisy data and labels thanks to the confidence information inherent to LF-MMI.

All the labels used in training this model are extracted from decode lattices, so biasing the lattices has a large impact. Here, the insertion reward improves the model by another 1.2% relative WER over the baseline.

4. Conclusions and future work

Building ASR systems for domains where data is scarce or absent is a challenging task. In this paper, we have shown the steps we have taken to build a system for a low resource language where the only transcribed speech present is from a different domain and bandwidth than the target data. We have also shown the validity of a method to adjust the generation of targets for semi-supervised LF-MMI.

In the future, we are interested to determine how the lattice biasing methods proposed interact with the strength of the acoustic and languages model used in generating the lattices. We are also interested in exploring methods to speed up the semi-supervised process, such as automated techniques for selecting the unsupervised data to use in adaptation before the decoding process.

5. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL) contract #FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

We thank Simon Vandieken for providing the language models used in this work, Alberto Abad for xVector training, and Anton Ragni and Mark Gales from Cambridge University for providing the multilingual bottleneck network.

6. References

[1] D. Smith, A. Sneddon, L. Ward, A. Duenser, J. Freyne, D. Silvera-Tawil, and A. Morgan, "Improving child speech disorder assessment by incorporating out-of-domain adult speech," in *Interspeech*, 2017, pp. 2690–2694.

[2] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *IEEE ASRU*, 2017, pp. 279–286.

[3] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *ICLR*, 2013, pp. 1–9.

[4] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[5] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–129, 2002.

[6] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *IEEE ICASSP*, 2004.

[7] P. Bell and S. Renals, "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *IEEE ASRU*, 2015, pp. 675–680.

[8] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[9] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *IEEE ASRU*, 2013, pp. 267–272.

[10] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *IEEE ICASSP*, 2006.

[11] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *IEEE ICASSP*, 2011, pp. 4656–4659.

[12] M. L. Seltzer and A. Acero, "Training wideband acoustic models in the cepstral domain using mixed-bandwidth training data for speech recognition," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, p. 1087, 2011.

[13] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *IEEE SLT*, 2012, pp. 131–136.

[14] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *IEEE ICASSP*, vol. 3, 2000, pp. 1843–1846.

[15] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Interspeech*, 2015.

[16] P. S. Nidadavolu, C.-I. Lai, J. Villalba, and N. Dehak, "Investigation on bandwidth extension for speaker recognition," in *Interspeech*, 2018, pp. 1111–1115.

[17] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *IEEE ICASSP*, 2018, pp. 4844–4848.

[18] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models," in *IEEE SLT*, 2018, pp. 250–257.

[19] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiat, S. Kombrink, P. Motlicek, Y. Qian, K. Riedhammer, K. Vesely, and N. T. Vu, "Generating exact lattices in the WFST framework," in *IEEE ICASSP*, 2012, pp. 4213–4216.

[20] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Pearson London, 2014, ch. 9.6, pp. 348–358.

[21] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, 2018, pp. 3743–3747.

[22] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," sep 2016, pp. 2751–2755. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0595.html http://www.danielpovey.com/files/2016_interspeech_mmi.pdf

[23] A. Ragni and M. Gales, "Automatic speech recognition system development in the "wild"," in *Interspeech*, 2018, pp. 2217–2221.

[24] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[25] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE ASRU*, 2013.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE ICASSP*, 2018.

[27] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *IEEE SLT*, 2012, pp. 336–341.